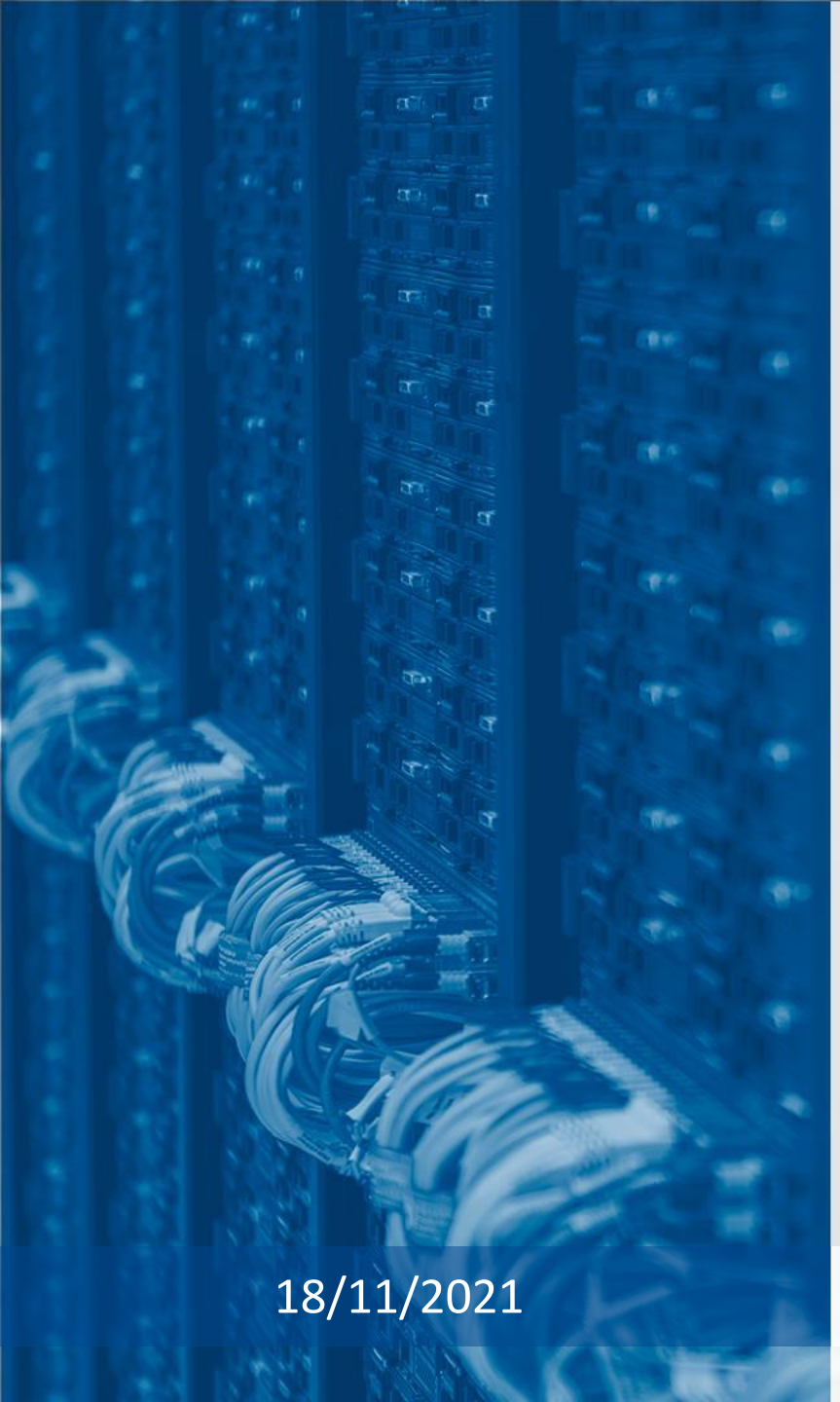


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

XIOS BOG

Xavier Yepes-Arbós
Miguel Castrillo

18/11/2021

EC-Earth Consortium General Assembly

Index

1. Context


BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Introduction

- The increase in the forecast capability of Numerical Weather Prediction (NWP) and climate modelling is strongly linked to the spatial resolution to solve more complex problems.
- This requires a large demand of computing power and it might generate a massive volume of model output which implies:
  - Data must be efficiently written into the storage system.
  - No more offline post-processing is affordable due to the size of the “raw” data.
  - A high cost of storage systems due to the huge data size.
- In this context, the improvement of the computational efficiency of NWP and climate models will be mandatory.

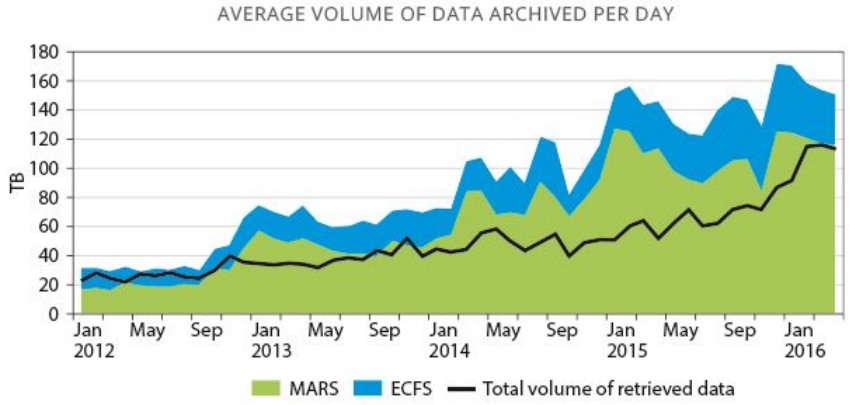
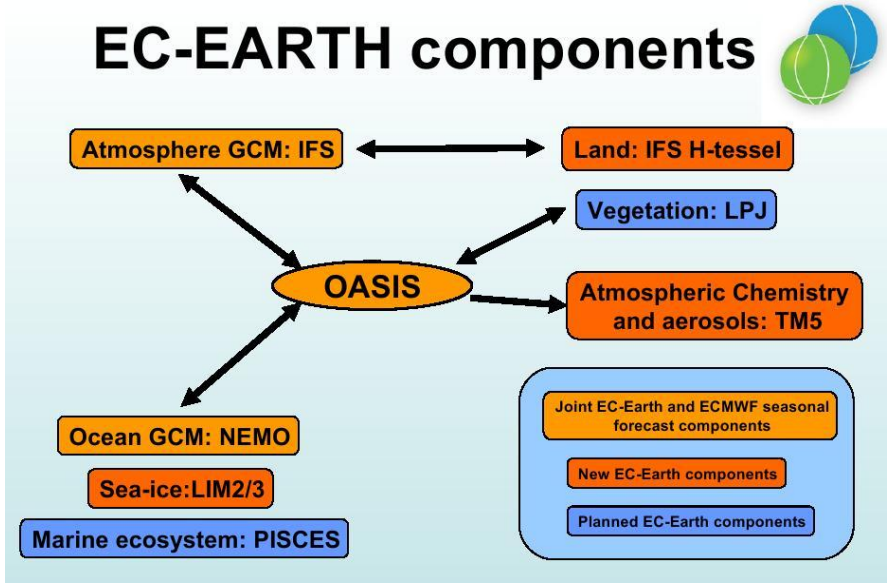


Figure source: ECMWF

EC-Earth overview

- Not much attention had been paid on **improving I/O** of Earth system models because it did not use to be an issue.
- EC-Earth 3 **couples** IFS CY36R4 to NEMO 3.6 and other Earth system components, using OASIS3-MCT.
- This IFS version of EC-Earth uses a **sequential I/O scheme**, which is **not scalable** for high resolution grids, and even less, for future exascale machines.
- EC-Earth was used to run T511L91-ORCA025L75 experiments under the H2020 PRIMAVERA project. Experiments were designed to output lots of fields, causing a considerable slowdown in the EC-Earth execution time (IFS I/O represented about **30%** of the total execution time).

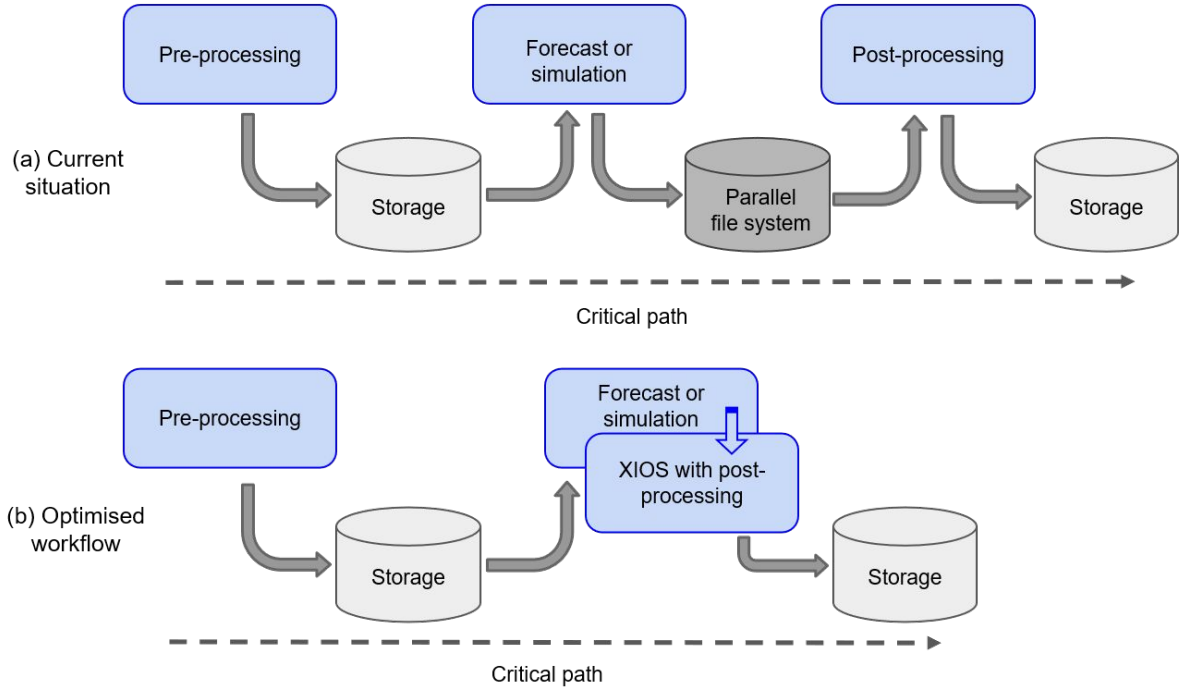


Objective

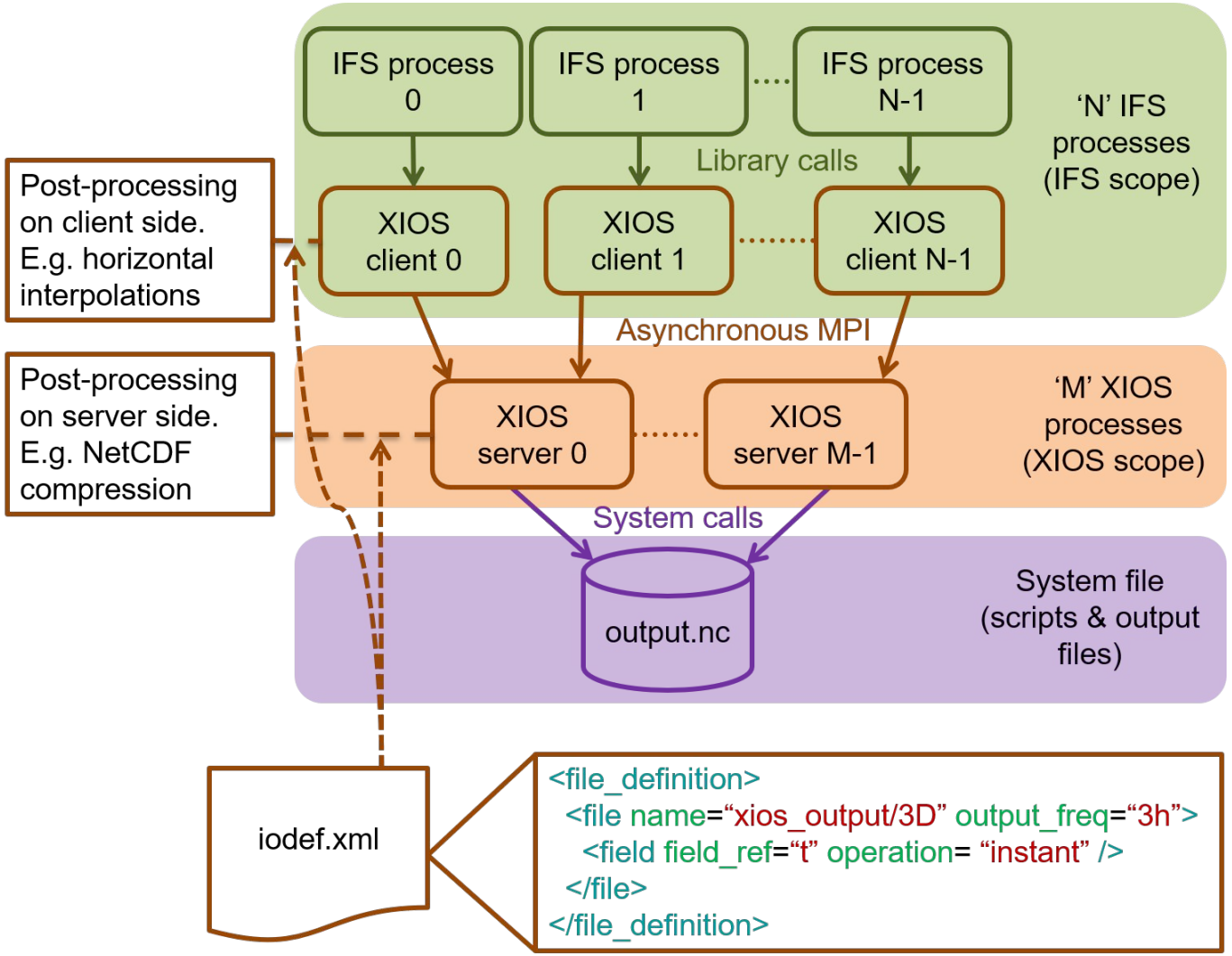
- Taking advantage that NEMO is already outputting data through XIOS, we chose to **integrate XIOS** into **IFS** as well.
- The XML Input/Output Server (XIOS) is an **asynchronous** MPI parallel I/O server developed by the Institute Pierre Simon Laplace (IPSL).
- The use of XIOS has the objective of improving the computational **performance** and **efficiency** of IFS (by extension EC-Earth), and thus, reduce the execution time.
- Moreover, it has a series of **additional benefits** (next slide).

How will EC-Earth benefit from XIOS?

- XIOS is a I/O tool widely used in the climate community because of these **features**:
  - Output files in **netCDF** format.
  - Written data is **CMIP-compliant** (CMORized).
  - It is able to post-process data inline to generate **diagnostics**.
- Future EC-Earth workflow:
  - **Critical path** will be **shortened** by concurrently running **post-processing** with the simulation.
  - **Simpler workflows** are less prone to have configuration errors.
  - **Lighter ece2cmor**. Move resources to dr2xml.
- Data **compression**.

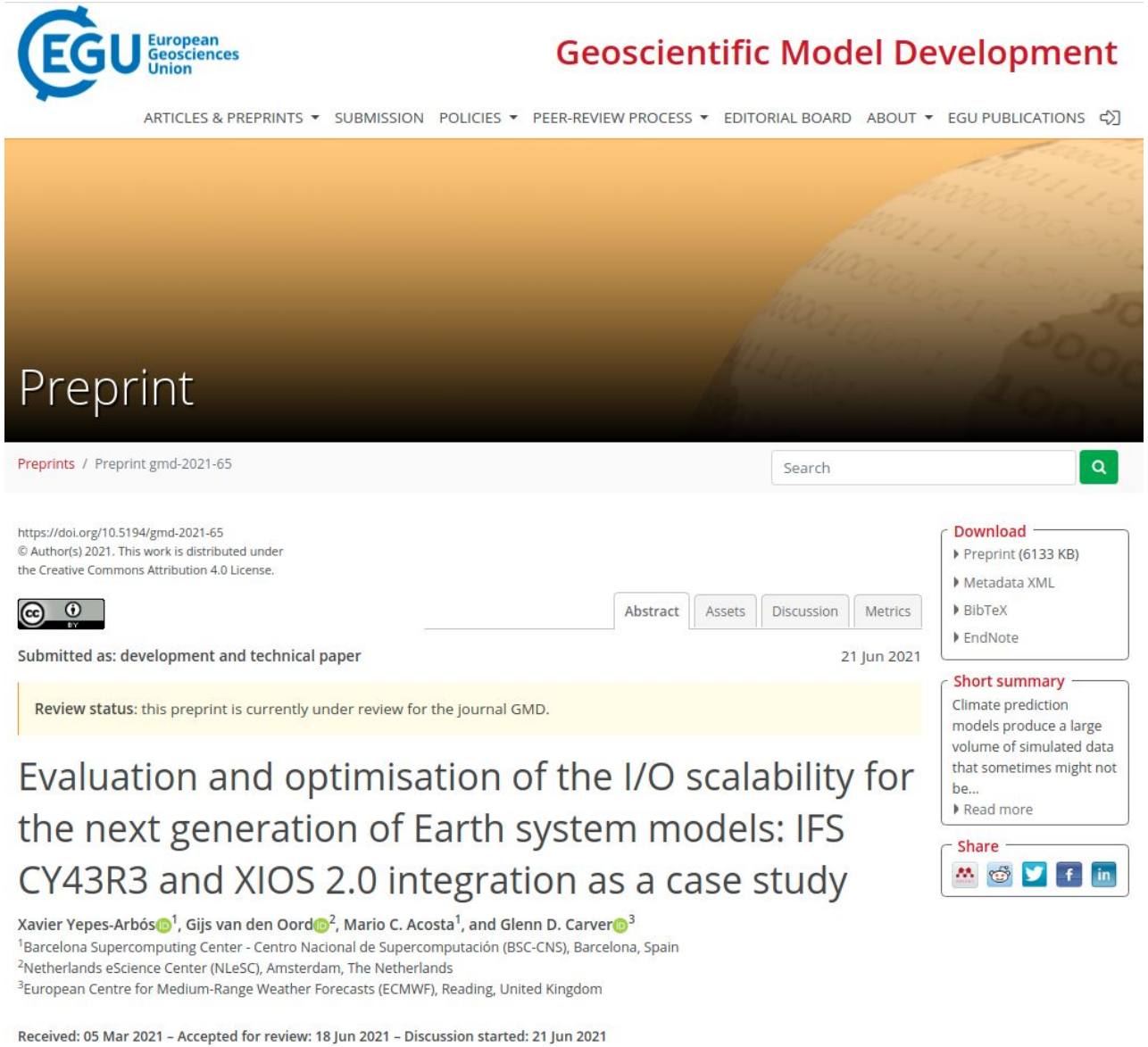


IFS-XIOS integration scheme



IFS-XIOS integration summary

- Scientific highlights:
  - Both **grid-point** and **spectral fields** (transformed to grid-point space using TRANS) are supported.
  - All **surface** and **3D** fields can be output.
  - Different **vertical levels** are available: model, pressure, theta and PV levels.
  - **No** longer needed to set up the **FullPos namelist** (NAMFPC).
  - FullPos **spectral fitting** is available.
  - Physical tendencies and fluxes output (**PEXTRA fields**) are also supported.
- Computational performance highlights:
  - In-depth benchmarking: **small** data output **overhead** with enough computational resources.



EGU European Geosciences Union

Geoscientific Model Development

ARTICLES & PREPRINTS · SUBMISSION · POLICIES · PEER-REVIEW PROCESS · EDITORIAL BOARD · ABOUT · EGU PUBLICATIONS

Preprint

Preprints / Preprint gmd-2021-65

Search

https://doi.org/10.5194/gmd-2021-65

Abstract · Assets · Discussion · Metrics

Submitted as: development and technical paper

21 Jun 2021

Review status: this preprint is currently under review for the journal GMD.

Evaluation and optimisation of the I/O scalability for the next generation of Earth system models: IFS CY43R3 and XIOS 2.0 integration as a case study

Xavier Yepes-Arbós[1], Gijs van den Oord[2], Mario C. Acosta[1], and Glenn D. Carver[3]

[1]Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS), Barcelona, Spain
[2]Netherlands eScience Center (NLeSC), Amsterdam, The Netherlands
[3]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, United Kingdom

Received: 05 Mar 2021 – Accepted for review: 18 Jun 2021 – Discussion started: 21 Jun 2021

Download
- Preprint (6133 KB)
- Metadata XML
- BibTeX
- EndNote

Short summary
Climate prediction models produce a large volume of simulated data that sometimes might not be...
Read more

Share

https://doi.org/10.5194/gmd-2021-65

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EC-Earth 4 and OpenIFS

- The new **EC-Earth 4** uses **OpenIFS 43R3** instead of IFS.
- OpenIFS is **derived** from IFS, which has the same forecast capability but the data assimilation functionality has been removed.
- To make use of XIOS in the atmospheric component of EC-Earth 4, the **IFS-XIOS** integration has been **ported** to OpenIFS 43R3.
- Aside from coupling OpenIFS 43R3 and NEMO 4 using OASIS3-MCT, SMHI has also configured EC-Earth 4 to allow both **OpenIFS** and **NEMO** writing their outputs with XIOS at the same time.

2. Usability

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Does it change how XIOS is used in EC-Earth?

- In XIOS terminology, each component of a coupled system that uses XIOS has its own **context**.
- A context is **self-contained**: grids, fields, files, etc.
- In general, contexts are **independent** of each other. However, it is possible to output diagnostics that combine fields from different contexts.
- NEMO XML files are configured as usual.
- The following slides will focus on the XIOS features that are **particular** to **OpenIFS**.
- A complete OpenIFS-XIOS user guide is available in this ECMWF webpage.

XIOS XML files for OpenIFS

- **iodef.xml**: common to both OpenIFS and NEMO to control basic XIOS parameters.
- **context_oifs.xml**: it contains FullPos parameters to tune the vertical interpolations.
- **axis_def_oifs.xml**: it defines the different types of vertical levels available.
- **domain_def_oifs.xml**: it defines the different domains (or types of grids) available.
- **grid_def_oifs.xml**: it defines the grids (XIOS terminology) available used to map fields.
- **field_def_oifs.xml**: it defines all the available fields to be output.
- **file_def_oifs.xml**: it defines NetCDF files to be written.

FullPos parameters

- It is possible to control some FullPos variables.
- FullPos spectral fitting:
  - NFITP
  - NFITT
  - NFITV
- Other types of variables:
  - NFPCLI
  - LFPQ
  - LTRACEFP
  - RFPCORR

```
<variable_group id="spectral_fitting">
  <variable id="nfitp" name="NFITP" type="int"> 2 </variable>
  <variable id="nfitt" name="NFITT" type="int"> 2 </variable>
  <variable id="nfitv" name="NFITV" type="int"> 2 </variable>
</variable_group>

<variable_group id="fullpos_other">
  <variable id="nfpcli"   name="NFPCLI"   type="int"   > 0       </variable>
  <variable id="lfpq"     name="LFPQ"     type="bool"  > false   </variable>
  <variable id="ltracefp" name="LTRACEFP" type="bool"  > false   </variable>
  <variable id="rfpcorr"  name="RFPCORR"  type="double"> 60000.0 </variable>
</variable_group>
```

Define different regular Gaussian domains

- OpenIFS has **two default grids**:
  - Native reduced Gaussian grid.
  - Regular lat-lon grid with 256 latitudes and 512 longitudes.
- It is possible to change the sizes of the regular grid or even to declare more than one regular grid. There are two key **attributes**:
  - 'ni_glo': the number of longitudes.
  - 'nj_glo': the number of latitudes.
- It is necessary to specify '<generate_rectilinear_domain />' to indicate that a horizontal interpolation is needed.

```
<domain_group id="regular_domains" type="rectilinear" >
  <domain id="regular" long_name="regular grid" ni_glo="512" nj_glo="256" >
    <generate_rectilinear_domain />
    <interpolate_domain write_weight="true" />
  </domain>
</domain_group>
```

Enable regular lat-lon grid output

- To output a field in a regular lat-lon grid it is necessary to use the attribute 'grid_ref'.
- For example, to output the 3D temperature field in a 256x512 lat-lon grid and model levels, you can use this XML code:

```
<field field_ref="t"    name="t"    grid_ref="regular_ml" freq_op="6h" operation="instant" />
```

where 'regular_ml' is defined in grid_def_ifs.xml as follows:

```
<grid id="regular_ml" description="3D interpolated regular grid with hybrid model levels" >
  <domain domain_ref="regular" />
  <axis axis_ref="model_levels" />
</grid>
```

Understanding output frequency, sampling frequency, NFRHIS and NFRPOS

- It is necessary to **distinguish** between the scope of attributes
  - 'output_freq' and 'freq_op' -> XIOS
  - 'NFRHIS' and 'NFRPOS' -> FullPos.
- **'output_freq':** it controls the **writing frequency** of a netCDF file.
- **'freq_op':** it controls the **sampling frequency**, this is, the frequency of sending data from the model (OpenIFS) to XIOS.
- **'NFRHIS'** & **'NFRPOS':**
  - They control the **post-processing frequency** of FullPos and the TRANS package to transform spectral fields to grid-point fields.
  - Must be set up taking the **greatest common divisor** (gcd) of all 'freq_op' values defined.
  - Note that **always** 'NFRHIS' = 'NFRPOS'.

Understanding output frequency, sampling frequency, NFRHIS and NFRPOS

Example: you want to output these three variables:

- 't': output every 6 hours the average of hourly data. You should set up 'output_freq=6h' and 'freq_op=1h'.
- 'u': output every 12 hours the maximum of 3 hourly data. You should set up 'output_freq=12h' and 'freq_op=3h'.
- 'cc': output every 6 hours instant data. You should set up 'output_freq=6h' and 'freq_op=6h'.
- 'NFRHIS' and 'NFRPOS' must be set up taking the gcd of 'freq_op' values across all fields ('t', 'u' and 'cc'), which is 1h. If the time step duration is 900 seconds for instance, then 'NFRHIS' and 'NFRPOS' should be set up to 4 (1h).

- It is also possible to specify these two FullPos variables in hours by using negative values: 'NFRHIS' = 'NFRPOS' = '-1'

Not correctly setting freq_op, NFRHIS and NFRPOS

It is very important to correctly set up 'NFRHIS' and 'NFRPOS' to produce correct data and do not waste computational resources:

- If 'NFRHIS' and 'NFRPOS' use a **smaller value** than the gcd, FullPos would be called **unnecessarily** (post-processed data will not be sent to XIOS as it is not required). This increases the computational cost with no gains in the accuracy of the results.
- If 'NFRHIS' and 'NFRPOS' use a **bigger value** than the gcd, FullPos would **not be called enough times** (post-processed data will not be correct).

Optimizations for sending data from OpenIFS processes to XIOS servers

- There are two available optimizations that might be useful to improve the execution time under some circumstances. It is not possible to predict in which conditions (many factors), so it is necessary to test them.
- They are disabled by default, but can be enabled in context_ifs.xml:
  - **'LOPT_SEND':** it enables a mechanism to change where data is sent from XIOS clients (OpenIFS processes) to XIOS servers to truly **overlap** OpenIFS computations with XIOS communications.
  - **'LSINGLE_PREC_SEND':** it sends data from XIOS clients to XIOS servers in **single precision** (32 bits) instead of double precision (64 bits). This allows you to send half of the data to decongest the network.

```
<variable_group id="client_server_comm">
  <variable id="lopt_send"         name="LOPT_SEND"         type="bool"> false </variable>
  <variable id="lsingle_prec_send" name="LSINGLE_PREC_SEND" type="bool"> false </variable>
</variable_group>
```

3. Scalability

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Computational performance considerations

One file vs. multiple file mode

- **'one_file'** mode has a limited computational efficiency as it **does not scale well** when outputting a large volume of data for **high resolution** configurations.
- **'multiple_file'** mode achieves a good computational efficiency as it **scales** with many resources.
  However, each XIOS server writes its own netCDF file, so **output data is splitted** between all these partial files.
  - It is necessary to study if there is any existing tool capable of efficiently combining these partial netCDF files into a single one (in OpenIFS).
  - If it does not exist, it would be necessary to develop such a tool, like in NEMO.

Computational performance considerations

Horizontal interpolations

- This kind of operation is **very expensive** depending on different parameters (number of fields, size of the fields, frequency, etc), so it can have a large impact on the total execution time.
- Why? Because interpolations are performed on the **client side** of XIOS, so OpenIFS processes have to first compute these interpolations before resuming the time stepping.
- Recommendation: use horizontal interpolations when it is strictly necessary and not systematically.

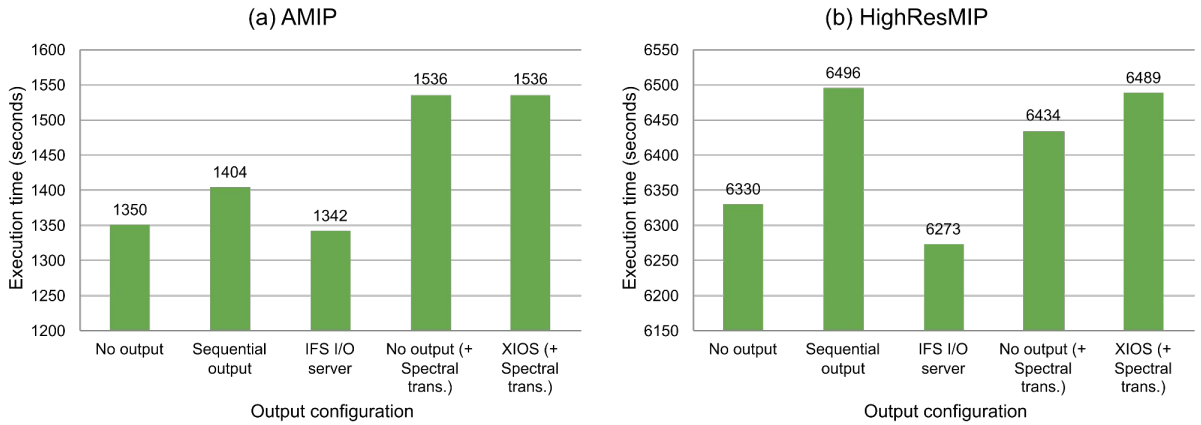
Computational performance considerations

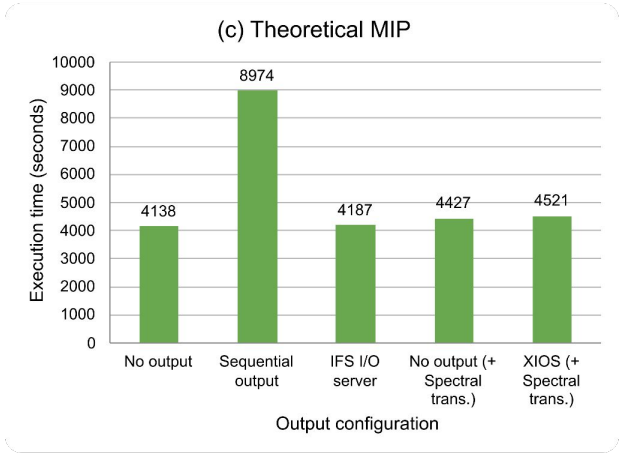
Lustre filesystem

- The **Lustre** filesystem stores a file in one or more Object Storage Target (OST) devices.
- If OpenIFS is run on a cluster that uses Lustre it is important to pay attention to the **striping**, which allows to divide a file into chunks that are stored in different OSTs.
  - When using the 'one_file' mode, it is important to set up a striping for each netCDF at least as equal as to the number of XIOS servers.
  - This allows each XIOS server to write into a different OST, which prevents to affect the performance of the whole system.

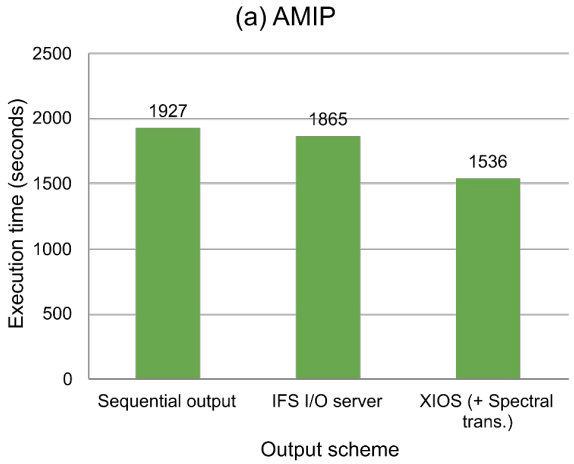
Computational performance of IFS-XIOS

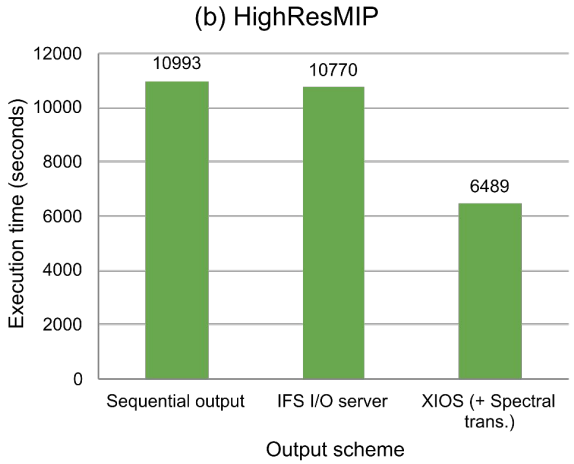
Output schemes comparison

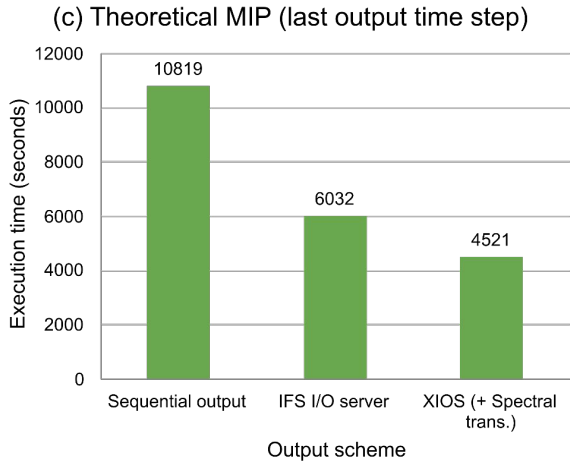




Output schemes comparison including post-processing





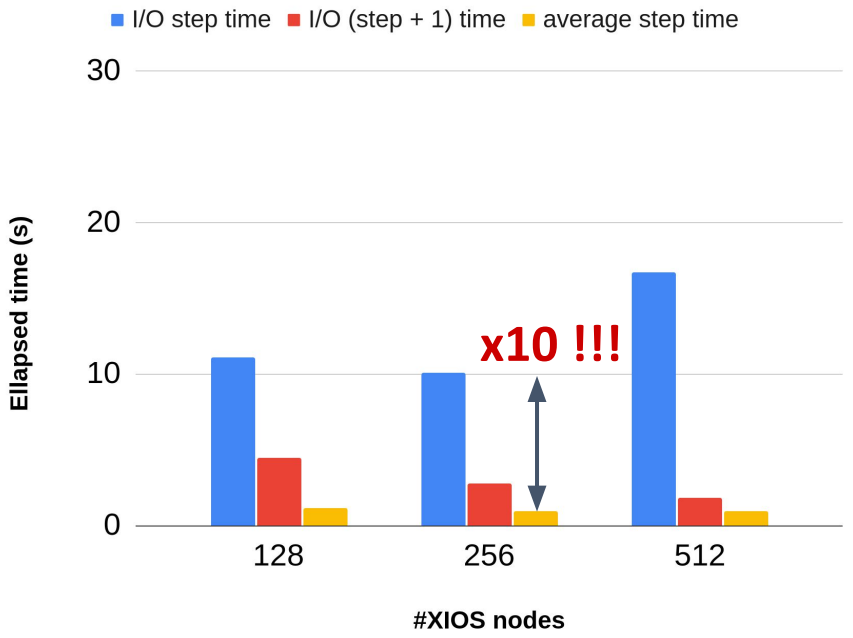


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

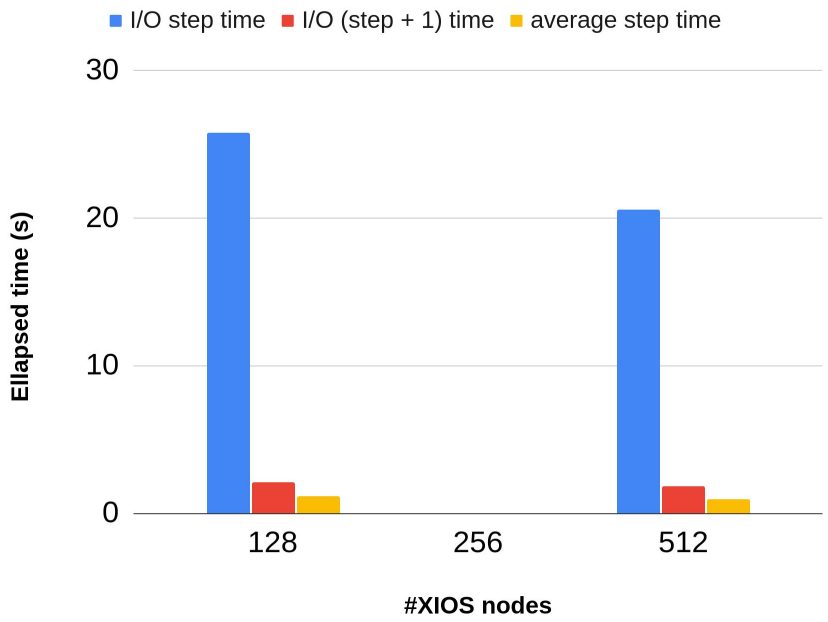
Scaling NEMO4-ORCA36 (with I/O)

- Measuring the **duration** of I/O steps (every 120th steps in this case)
- **Multiple-file** mode (one-file mode takes **minutes**)
- **High memory** needs
- XIOS not scalable

Performance mode
(large buffer)



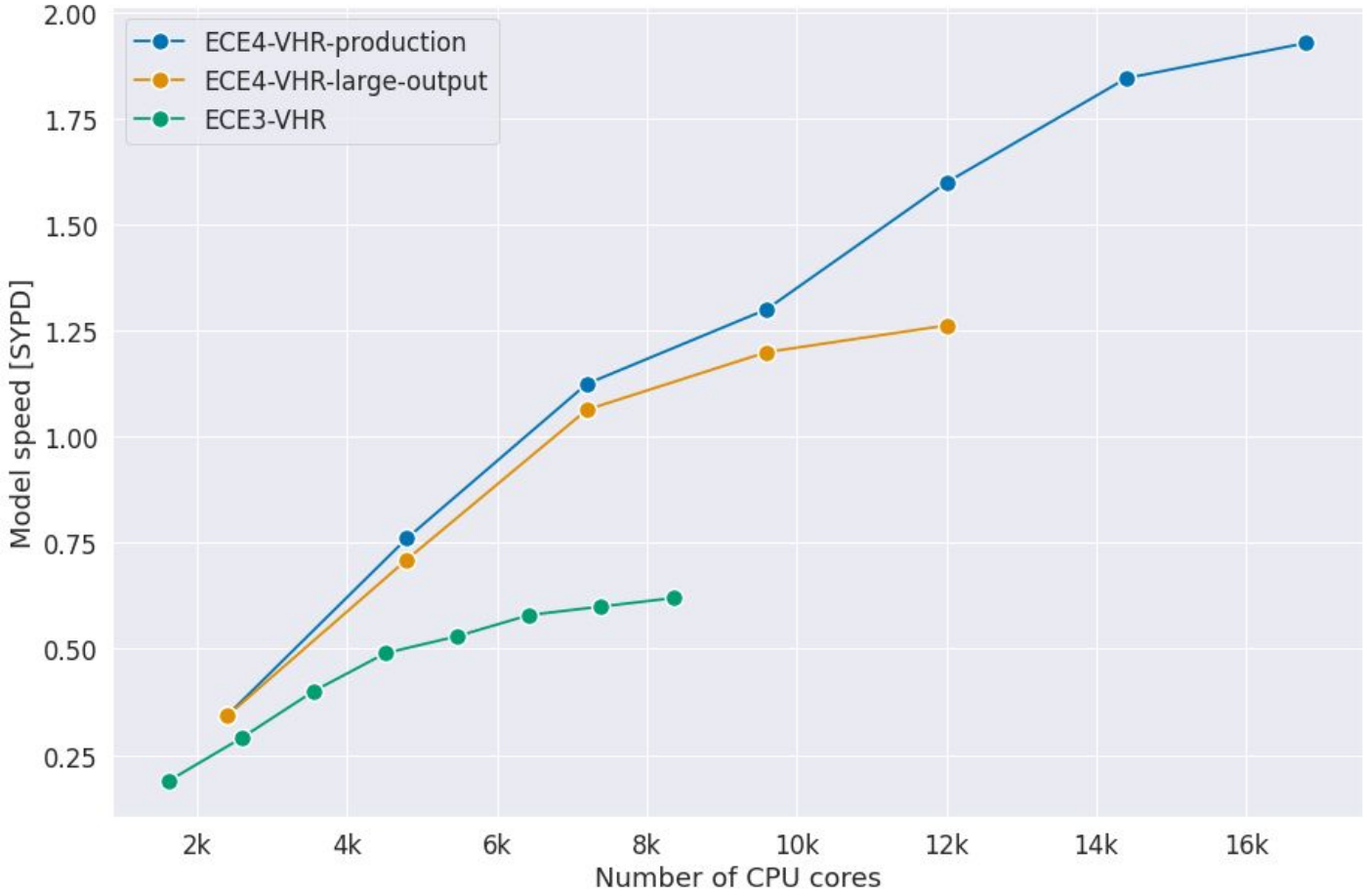
Memory mode
(large buffer)



NEMO running on 512 nodes
1-hour 3D output (340 GB per hour)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Tco639-ORCA12 in MareNostrum 4



ECE3: TL1279-ORCA12

ATM: 360s, **OCE: 360s**, ICE: 720s, **CPL: 720s**

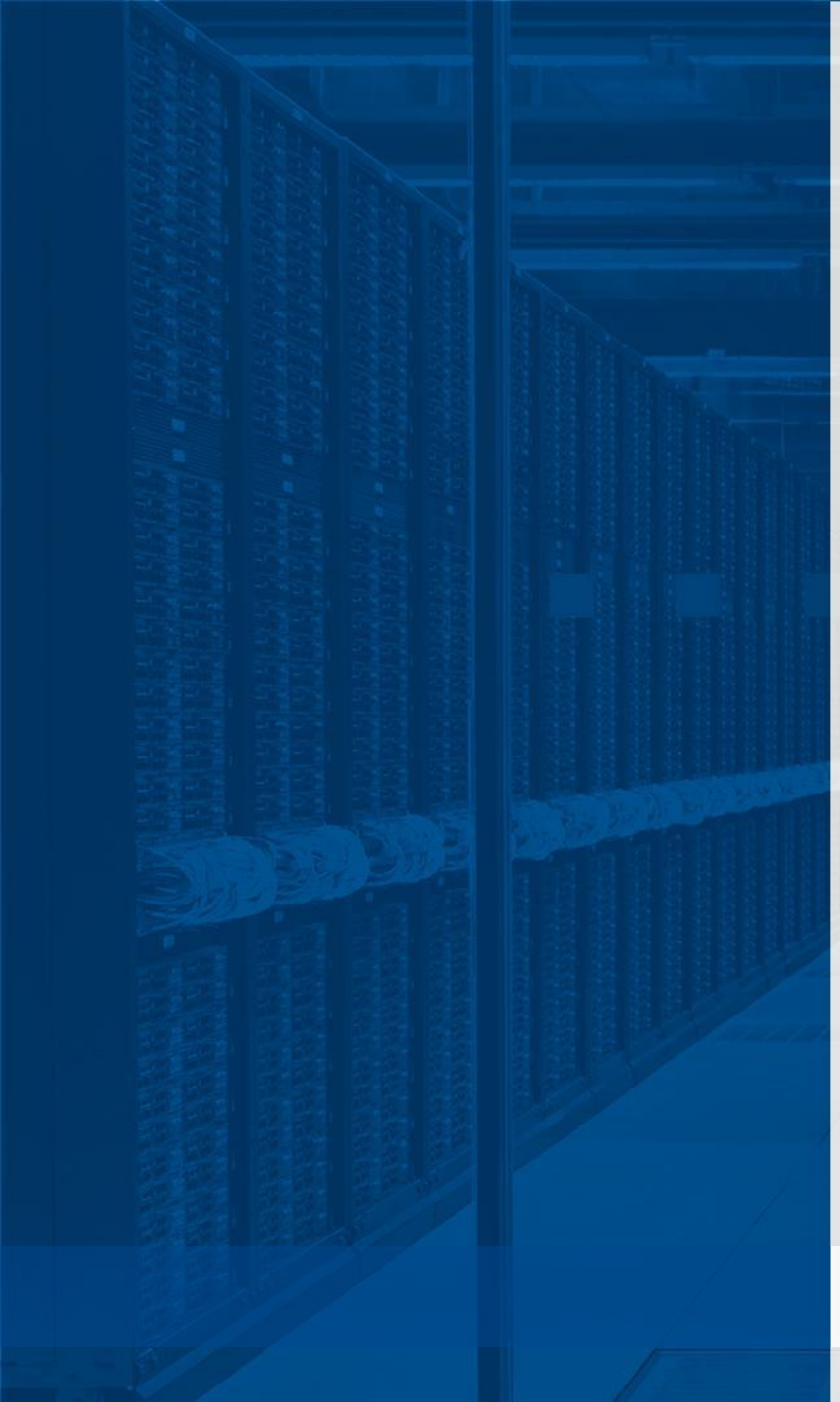
ECE4: Tco639-ORCA12

ATM: 360s, **OCE: 240s,** ICE: 720s, **CPL: 3600s**

- **Production: Monthly** output (6-hourly and daily averages)
- **Large output: 3-hourly** output

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

100 year experiment, **~16M** computing hours & **68** days



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thank you

xavier.yepes@bsc.es miguel.castrillo@bsc.es